Web on the Wall Reloaded: Implementation, Replication and Refinement of User-Defined Interaction Sets

Michael Nebeling, Alexander Huber, David Ott and Moira C. Norrie
Department of Computer Science, ETH Zurich
{ nebeling, norrie }@inf.ethz.ch

ABSTRACT

System design using novel forms of interaction is commonly argued to be best driven by user-driven elicitation studies. This paper describes the challenges faced, and the lessons learned, in replicating Morris's Web on the Wall guessability study [6] which used Wizard of Oz to elicit multimodal interactions around Kinect. Our replication involved three steps. First, based on Morris's study, we developed a system, *Kinect Browser*, that supports 10 common browser functions using popular gestures and speech commands. Second, we developed custom experiment software for recording and analysing multimodal interactions using Kinect. Third, we conducted a study based on Morris's design. However, after first using Wizard of Oz, Kinect Browser was used in a second elicitation task, allowing us to analyse and compare the differences between the two methods. Our study demonstrates the effects of using mixed-initiative elicitation with significant differences to user-driven elicitation without system dialogue. Given the recent proliferation of guessability studies, our work extends the methodology to obtain reproducible and implementable user-defined interaction sets.

Author Keywords

guessability studies; mixed-initiative design; user-defined multimodal interaction sets.

ACM Classification Keywords

H.5.2. User Interfaces: Input devices and strategies

INTRODUCTION

Within the larger HCI community, there is currently a lot of interest in user-defined gesture, speech and multimodal interaction sets and how they may inform the design of new systems. Given the recent proliferation of guessability studies, this paper raises two important issues. First, most studies thoroughly adopt Wobbrock et al.'s [14] methodology in order to obtain a user-defined interaction set, but without considering implementation issues or involving experts. Rather, it is common to completely remove system dialogue to focus on users and their interaction preferences. The majority of studies end with reporting a suitable interaction set based on computed agreement scores and consensus thresholds. However, important aspects for system design, such as conflicts and consistency of suggested interactions, as well as the design implications of supporting alternative interactions, are often not even considered. The second important issue we address is replication. Revisiting, replicating and reproducing HCI findings is increasingly valued within our community. Still, there are many issues that make it a challenge for our discipline [2]. In particular for guessability studies, there is a lack of guidelines and tools to support replication.

This paper describes our investigation around *Kinect Browser*—a new multimodal web browser that we designed to support 10 common browser functions using Kinect. Kinect Browser was developed based on Morris's Web on the Wall study [6] that focused on eliciting multimodal interactions by potential end-users using Wizard of Oz. However, her guessability study alone was not sufficient to design our system. Despite her thorough investigation, we as system designers struggled to identify a conflict-free and consistent interaction set from her findings that would be feasible to implement. In addition, there were other fundamental questions concerning the implementation using Kinect that her study did not address. We therefore decided to replicate and extend Morris's Web on the Wall study. In addition to user-driven elicitation using Wizard of Oz, we investigate the impact of using a system both for recognition of suggested interactions and execution of associated functions. Compared to Morris's study, ours produced similar results, but also revealed potential issues of the methodology, demonstrating that it is vital to include system dialogue even in the early stages of design.

This paper is mainly about the design process of Kinect Browser. Being a functional prototype of a Kinect-based web browser, it differs from previous research such as [4] that targeted developers with the goal of providing a JavaScript library for processing Kinect's data streams within the browser. While our system in itself offers a valuable contribution, we see three primary contributions in our work. First, we add to the growing body of knowledge surrounding the optimal methodology for user-driven elicitation studies, such as those popularised by Wobbrock et al. [14]. Our elicitation study is the first to pay attention to what the system can do while asking users what they would like to do. Second, our paper also provides data about popular gestures and speech commands for a living room TV setting. The data was collected using new tools specifically developed for recording and post-hoc analysis of Kinect's depth, video and audio streams. We

ITS 2014, November 16–19, 2014, Dresden, Germany.

http://dx.doi.org/10.1145/2669485.2669497

Source	Methods (Size)	Focus; Main Findings/Contributions
Wobbrock et al. [14]	Think-aloud + Video + Logs ($N = 20$)	**user-defined gestures;** 27 referents with author-rated complexity scores; user agreement scores; taxonomy of surface gestures
Nacenta et al. [9]	Custom experiment software ($N_1 = 6$, $N_2 = 9$, $N_3 = 12$)	**gesture memorability;** user-defined better than pre-designed or random gestures; differences due to association errors rather than gesture form errors
Oh et al. [11]	Think-aloud + Video + Logs + \$N recogniser ($N = 20$)	**personal gestures;** focus on familiar rather than novel gestures; constrained by gesture recognition issues; proposal for mixed-initiative gesture definition
Morris [6]	Wizard of Oz ($N = 25$; 11 pairs, 1 triad)	**multimodal interactions;** 15 referents; mostly speech, rarely multimodal—instead gesture synonyms; max-consensus and consensus-distinct ratio

Table 1. Previous user-driven elicitation studies that influenced the design of the one presented in this paper.

share both the code and the data with the community. We consider this generally useful to those developing Kinect-based systems. Finally, the paper offers a contribution in the form of replication of prior work from our research community, namely Morris's Web on the Wall study [6]. Understanding which aspects of her findings did and did not replicate in our study is interesting for generalisability. In particular, we discuss using Wizard of Oz for user-driven elicitation, focusing on issues of reproducibility and implications for design.

RELATED WORK

One of the most important design challenges for Kinect Browser is the rich design space when it comes to gesture, speech and multimodal interaction sets using Kinect. In building Kinect Browser, we felt that there is currently still a lack of guidelines and tools to address this issue in a way that is both reproducible and technically sound. Table 1 gives an overview of the studies that formed the basis of our own investigation, summarising their methods and main findings.

Wobbrock et al. [14] were the first to suggest a study design that focuses on the user by excluding the system dialogue from the investigation. Their elicitation task was based on showing the *effect* of a gesture, and then prompting users for the *cause* by asking them to perform suitable one or two-hand gestures. As the suggestions were quite diverse, they provide a first agreement metric and user scores together with a taxonomy of surface gestures. The other studies are based on this methodology for elicitation tasks. With the exception of Morris's [6], most studies are focused on gestural interaction and commonly use the think-aloud protocol and video analysis in combination with user logs. Also our study adopts a similar design, but additionally uses Kinect for both data collection and analysis. While our intention is to support reproducibility and help build a corpus of multimodal interactions using Kinect by sharing our data with the community, it is often not clearly described for previous studies whether and how they made use of any collected data, even when touch surfaces [8], mobile phone accelerometers [12], or Kinect sensors [3] were used as additional recording tools.

Previous studies have shown several advantages of user-defined gesture sets. For example, Nacenta et al. [9] conducted three experiments on gesture memorability, showing higher rates for user-defined gestures compared to pre-designed and stock gesture sets. Interestingly, the relatively high differences in recall rates were due to association errors rather than gesture form errors. In their study, user-defined gesture sets were also preferred and found easier to learn and remember. Although not formally assessed, memorability also played a role in our study, as users had to remember and repeatedly perform custom multimodal interactions.

The study by Oh et al. [11] focused on gesture customisation. Even though participants were encouraged to iteratively personalise gestures, they generally focused on the familiar rather than creating novel gestures. Post-hoc analysis showed relatively high recognition rates for a multistroke extension of \$1 recogniser [15], but users' design choices were sometimes biased by misconceptions about the recogniser's abilities. To improve the gesture creation process, the authors suggest a mixed-initiative approach in which the process is monitored and, once an ambiguous gesture has been recognised, users would be prompted for, or automatically suggested, modifications to improve the gesture. In our investigation on Kinect Browser, we implemented such a mixed-initiative approach by adding a task in which users defined, tested and *refined* their multimodal interactions while using Kinect.

Our design decisions for Kinect Browser were based on the observations made by Morris [6]. She adopted the study design proposed by Wobbrock et al. [14] for eliciting speech commands and multimodal interactions in addition to gestures for web browsing on a living room TV. Along the scenario of planning a shared weekend activity, Wizard of Oz was employed to demonstrate 15 common web browser functions, showing the effect of each function and asking users to define an interaction that may have caused it. A key aspect addressed in her paper is how to measure consensus among participants, which becomes an issue due to the many degrees of freedom in which multimodal interactions can be defined. Specifically, the paper provides two new agreement metrics: *max-consensus* and the *consensus-distinct ratio*. The first results in the percentage of participants that suggested the most popular interaction, while the latter represents the number of interactions that achieved a consensus threshold of two in proportion to the total number of proposed interactions.

In her study, both gesture and speech seemed effective. Speech was most commonly suggested, followed by gesture, while multimodal suggestions composed of gesture *and* speech were rare. Rather, one of the key observations was that the same function was triggered by multiple independent interactions of the same user, instead suggesting *multimodal synonyms* using either gesture *or* speech to invoke the same action. This was explained by shortcomings of both modalities. First, gesture input had perceived drawbacks related to

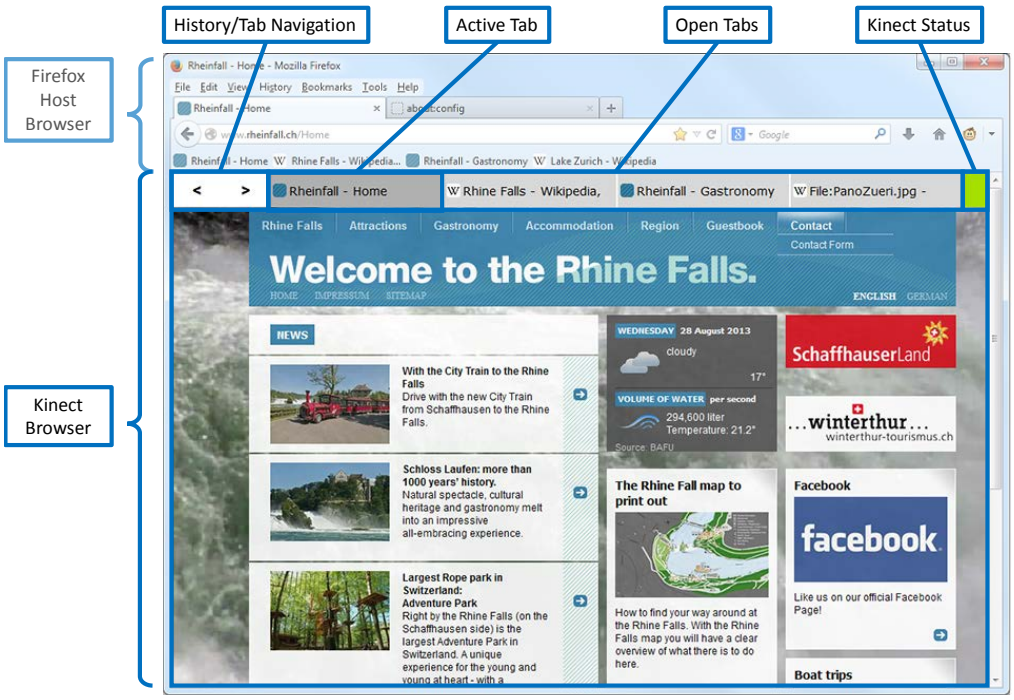

Figure 1. Kinect Browser was implemented as a "browser-in-browser" interface, but always operated in fullscreen so that the actual host browser, Firefox, was not visible to participants.

ergonomics and exertion. But also speech interaction was often regarded as potentially disturbing to other members of the household, while ambient noise and conversations might impact speech recognition.

Most of her participants assumed that their hands would be tracked similar to using a computer mouse, and basic interactions such as clicking links, scrolling or zooming were based on mouse and touch interaction paradigms. In addition to typical click and press gestures for confirming selections, also holding positions for a certain amount of time, i.e. dwell, was frequently suggested. We considered all these proposals to be suitable for Kinect Browser. While her participants were free to sit, stand, or otherwise move about, all of them remained seated for the elicitation. Likewise, Kinect Browser is currently optimised for seated mode, and so are the gestures which are currently limited to the upper body.

However, for several reasons, it was not straightforward to build on Morris's study and learn from her suggested design implications. First, her study did not actually involve the use of Kinect and consider essential browser functions such as scrolling and zooming. Second, not all results were conclusive. For example, the elicited interaction set is not conflict-free and still requires adjustments to individual user preferences. Third, some of her design decisions may have biased some of the results. For example, users were influenced by their previous experience with the existing desktop browser that was used for the Wizard of Oz parts. Finally, users were recruited in pairs and worked as a team to propose interactions, which may impact the study protocol and ultimately exclude individual use scenarios. These issues together motivated us to conduct a study of our own.

KINECT BROWSER

Although specifically designed for our elicitation study, Kinect Browser is a functional prototype that can be used for browsing most existing sites on the web without any modifications. The core of Kinect Browser consists of three components: a *browser-in-browser interface* supporting common browser functions (Figure 1), a *gesture and speech recogniser* based on Kinect, and a *multimodal configuration* mapping gestures and/or speech commands to browser functions.

The interface of Kinect Browser is shown in Figure 1. Similar to existing browsers, Kinect Browser leaves most of the screen real estate to the web site, but adds a special function area on top. Our function area consists of a set of icon buttons that can be associated with browser functions, e.g. for going back and forward in the history or to the previous and next tab. The remainder of the function area is filled with tabs. The area in the top-right corner gives feedback about the connection status, and can also be used to connect to, or disconnect from, Kinect. Therefore, Kinect Browser's design is kept rather simple and does not resemble any specific web browser in order to reduce bias.

Our current implementation supports 10 browser functions: Click Link, Scroll, Zoom In/Out, Go Back/Forward, Select Tab, Next/Previous Tab, and Reload Page. This means that most of the 15 referents proposed in [6] are covered by our implementation. In addition, we also included support for scrolling and zooming as essential browser functions, and explicitly distinguished selecting a specific tab and switching to the next or previous tab. In further distinguishing Morris's Switch Tab referent, we wanted to raise potential conflicts with gestures and speech commands for going back and for-

ward in the navigation history, and study whether and how users would aim to resolve any conflicts. At the same time, we ignored advanced referents such as Open Link in Separate Tab, Select Region, Bookmark Page, Close Tab and others requiring text input. Text input using Kinect, as required for the Search Engine Query or Enter URL referents, is an important research issue of its own out of scope of this paper. In a first step, we aimed to support all essential browser functions required for navigating within and between pages to assess the basic browsing experience based on multimodal interactions.

Kinect Browser tracks both hands independently, showing two cursors for the left and right hand on the screen. Hand positions within a 50x50cm interaction window centred on the Kinect origin are mapped to screen coordinates. This setting may require calibration with regards to different screen and user characteristics (size, position, distance, etc.). Based on an iterative design process and inspired from Morris [6], we added support for 25 interactions: 9 gestures (dwell/grip/press, drag up/down, flick hand left/right, two-hand pinch open/close) and 16 speech commands ("scroll up/down", "zoom [in/out]", "[go] back/forward", "tab #", "next/previous tab", "refresh/reload [page]"). We complemented our recogniser with $1 [15] to support unistroke gestures, e.g. circle hand. This required a mapping from 3D space to 2D, which we currently achieve by only mapping the x and y-axes. Note that we did not add explicit recognition support for interactions comprising both gesture *and* speech, as these were rarely proposed in Morris's study.

To adapt to users' multimodal interaction preferences, Kinect Browser currently builds on a simple configuration tool in which variables can be toggled and set to the desired values.

STUDY DESIGN

Morris [6] used a three-part study based on *interviews* concerning possible scenarios for web browsing on a living room TV using Kinect, an *elicitation task*, and a *post-study questionnaire*. Our study was composed of five parts with two elicitation tasks followed by a task in which Kinect Browser was used individually and tested in its anticipated setting.

Questionnaires and Tasks

We used a *pre-study questionnaire* in the beginning of the study asking participants to rate likeliness of the scenarios and importance of browser interactions proposed by Morris, as well as collecting background information such as prior influences in terms of operating systems and browsers, knowledge of Kinect, participants' handedness, age and gender.

The actual elicitation was conducted in two parts. *Task 1* was based on the elicitation task of Morris using Wizard of Oz to collect interactions proposed by participants for the 10 browser functions supported in Kinect Browser. Like Morris, we told, and always reminded, participants that they could use any combination of gesture, speech or combined input, and that they should not worry about the capabilities of Kinect, as the experimenter would act as the Wizard of Oz and ensure that the system "reacted" properly to their interactions. As in previous user-driven elicitation studies, for each function, the experimenter stated the function name and demonstrated the *effect*, then asked the participant to suggest interactions as a *cause* of this function to be executed. We used Kinect for recording and tracking in the background, but similar to Wobbrock et al. [14], no feedback was provided and no interactions were triggered by the system. When participants demonstrated an interaction, the experimenter used the mouse and/or keyboard to produce the result of the browser function. As this task was partly designed to replicate Morris's study, we will use it later to compare and correlate the findings.

Figure 2. Study setup consisting of a Kinect and a large screen for Kinect Browser, as well as a camera for video recording.

Task 2 followed a similar design, but the goal was now to produce a personal configuration for Kinect Browser. In this task, Kinect Browser replaced the Wizard part and was step-wise adapted to react to preferred interactions. Again, participants elicited interactions for each browser function. When participants suggested multiple interactions or interactions that Kinect Browser was not able to recognise, the experimenter suggested modifications to the suggested interaction, or the closest interaction implemented by Kinect Browser. Participants were then asked whether they would still prefer the interaction they suggested initially, or would like to switch to the one available in the system. As participants suggested interactions for each browser function, the experimenter supplied the necessary settings to Kinect Browser, and asked participants to demonstrate the interaction and test whether the system reacted as expected.

The common scenario for both tasks was to plan a (shared) weekend activity, in our example a trip to the Rhine Falls, where participants were asked to look up facts on the Rhine Falls web site[1] as well as Wikipedia. All suggested interactions were scored according to preferences. Each task concluded with a *post-task qustionnaire* based on Morris's post-study questionnaire [6], collecting subjective feedback on both elicitation tasks, allowing us to check for differences.

Technical Setup

The setup used for the study is illustrated in Figure 2. A dual-monitor setup was used to show Kinect Browser on the participant's screen and all interactions as they were tracked by Kinect on the experimenter's screen. The participant's screen was a ca. 63" projection surface with 1024x768 resolution. The Kinect was set to seated mode and placed 3.5 metres

[1] http://www.rheinfall.ch/home

away in front of the couch where participants were seated at 5 metres distance to the screen. Morris's original study was conducted using a 63" TV at 1600x1200 pixels at 3.5 metres distance. We used the same distance between Kinect and participants in our study, but the distance to the screen was increased to counterbalance the lower resolution.

In addition to creating video recordings of participants' interactions, we used custom experiment software specifically developed for our multimodal elicitation study to be able to record skeletal tracking information from Kinect together with audio and video. We also added logging to Kinect Browser, allowing simultaneous recording of browser interactions and Kinect data. Log files can be loaded into the experiment software for visualising all tracked events on a timeline. After the study, we used the data to produce skeletal tracking statistics and correlated it to user feedback.

Participants

Like Morris, we recruited 25 participants in order to enable comparison in terms of the max-consensus and consensus-distinct metrics. The median age of our sample was 26.5 years; 6 were female, and all participants right handed. Rather than doing the study in pairs, we conducted individual, 1-hour sessions with all participants, but allowed for discussions of interaction proposals. While this carries the potential of altering the results, it was necessary for Task 2 where otherwise specific skeleton selection strategies and a co-browsing interface [7] would have been required.

Also in contrast to Morris's study, our participants had little to no experience with Kinect. The idea here was to reduce potential bias towards certain interactions based on conventions emerging in the gaming community, which was noted by Morris. The majority of participants stated that they use Chrome or Firefox on Windows or Mac, followed by Android, iOS and Linux with their respective web browsers, which we collected as other potential influences. Again as in Morris's study, our pre-study questionnaire indicated that most participants could identify with the scenario of choosing a movie to see, followed by finding and viewing photos online, and looking up trivia and facts (all on average above 3 on a 5-point Likert scale). Other scenarios proposed by Morris, such as using social media applications and research for work or school projects, scored much lower for our participants, which may indicate further differences between the two samples. Additional scenarios frequently suggested by our participants included trip planning, which was the scenario used for the remainder of the study, as well as watching videos on Youtube, or checking news web sites. The most important browser functions, as rated by participants, included going back and forward in the history (median 7 on a 7-point Likert scale), tabbed browsing, bookmarks, and search (all median 6). Other functions explored by Morris, but that we did not include, such as finding or selecting content within the page, seemed less important (median 5).

RESULTS

This section presents the results of our investigation. We start with our participants' preferred interactions. This is directly followed by feedback and observations from our two elicitation tasks as well as user feedback on Kinect Browser. Special emphasis will then be given to the post-task ratings in our study and how they compare to Morris's post-study ratings.

Referent	Interaction	#		
		Web on the Wall	Kinect Browser	
			Task 1	Task 2
Click Link	hand-as-mouse + click/grip	7	11	18
	hand-as-mouse + dwell	6	0	0
	hand-as-mouse + push/press	N/A	5	1
Scroll Page	arm out and move same	(0-2)	11	2
	arm out and move opposite		7	0
	grip and drag opposite	N/A	5	17
	grip and drag same		1	6
Zoom In	two-hand pinch	(1-2)	18	12
	"zoom"	1	0	0
	"zoom in"	0	3	9
Zoom Out	two-hand pinch	(1-2)	17	11
	"zoom out"	0	4	9
Go Back	"back"	7	3	5
	flick hand (arrow)	7	5	4
	flick hand (book)	4	10	9
	"go back"	N/A	4	7
Go Forward	"forward"	6	3	6
	flick hand (arrow)	5	5	4
	flick hand (book)	5	9	9
	"go forward"	N/A	5	6
Select Tab	click tab	7	14	13
	"tab <number>"	3	3	6
	"tab <title>"	N/A	6	2
Next Tab	"next tab"	4	2	8
	flick hand (book)	3	4	2
	select tab	N/A	10	12
Previous Tab	"previous tab"	N/A	2	7
	flick hand (book)		4	2
	select tab		9	10
Reload Page	"refresh", "refresh page"	9	4	5
	"reload", "reload page"	N/A	6	10
	move hand in spiral motion	3	8	9

Figure 3. Preferred mapping of browser functions to gesture/speech interactions with number of occurrences according to Morris's and our own study. For both studies, the highest-scoring referent(s) for each metric are indicated with grey shading (consensus threshold 3).

Preferred Browser Control Techniques

Figure 3 shows popular mappings based on Morris's consensus metrics. As in Morris's study, participants suggested a range of gesture-based and speech-based interactions, and there was no uniform interaction set. Prior to the study, also the authors scored their preferences based on their experience of designing the system. Some of their choices matched several popular mappings. Yet, none of our 25 participants, and also none of the authors, proposed the same configuration. Also nobody suggested the one that would result from associating only the most popular interaction with every browser function. Nevertheless, there is good correlation of our results with Morris's in terms of the most popular interactions.

With a total of 449 interactions in Task 1 and 458 interactions in Task 2, participants proposed an almost equal number of interactions in both tasks independent of the Wizard of

Referent	Task 1 (Gesture)	Task 1 (Speech)	Task 1 (Multi...)	Task 2 (Gesture)	Task 2 (Speech)	Task 2 (Multi...)
Click Link	60%	16%	24%	82%	9%	9%
Scroll Page	80%	20%	0%	79%	21%	0%
Zoom In	66%	34%	0%	50%	50%	0%
Zoom Out	67%	33%	0%	55%	45%	0%
Zoom In/Out	67%	33%	0%	52%	48%	0%
Go Back	62%	38%	0%	56%	44%	0%
Go Forward	63%	37%	0%	57%	43%	0%
Select Tab	60%	40%	0%	59%	41%	0%
Next Tab	69%	26%	5%	54%	46%	0%
Previous Tab	71%	23%	6%	54%	46%	0%
Switch Tab	66%	30%	3%	56%	44%	0%
Reload Page	56%	44%	0%	45%	55%	0%
Total	**65.26%**	**30.73%**	**4.01%**	**59.61%**	**39.30%**	**1.09%**

Figure 4. Proportions of 449 proposed interactions in Task 1 and 458 in Task 2 showing the overall preference for gesture for both tasks and a gradual transition to speech when actually using the Kinect in Task 2.

Oz or mixed-initiative elicitation method. Despite the large amount of proposed interactions, our current implementation of Kinect Browser already covered 691 (76%) of them. This means that more than 3/4 of the interaction proposals by users were fully supported and could already be tested. Not covered were the small amount of multimodal interactions and some, usually distinct, gestures and speech commands.

Interestingly, in Task 2, participants tended to refine their interactions or made slightly different proposals compared to Task 1. The majority of participants only learned in Task 2 about the capabilities and limitations of the Kinect sensor and many argued that this prompted them to adjust their interactions. However, the preferred interactions typically stayed the same. As we will discuss later, the main difference between the proposals in Task 1 and Task 2 is in the Scroll Page referent were participants commonly switched from the arm-out-and-move to the grip-and-drag interaction using their hands.

In terms of modality, the preferred interactions in Morris's study were composed of 56% speech, 41% gesture, and 3% multimodal commands. As shown in Figure 4, our participants most commonly suggested gesture (Task 1: 65%, Task2: 60%), then speech (Task 1: 31%, Task 2: 39%), and also rarely multimodal (Task 1: 4%, Task 2: 1%). Similar to Morris's study, participants noted several drawbacks for both modalities. In particular in view of scenarios such as listening to music, talking to a colleague on the phone, having friends over, etc., speech was considered inferior (P9: *"This has to work with gestures, or it's not gonna work at all for me"*).

Overall, there was the trend of using gesture for more frequent functions. One exception was P10 who advocated the assignment of voice commands for more frequent functions, and gesture for less frequently used functions, since speech input seemed "surprinsingly" robust. As in Morris's study, participants commonly defined *multimodal synonyms*, stating that the preferred interaction depends on the situation, and that they would like to have both gesture commands and speech commands available. Still, P21 noted, *"there should be the possibility to have an only gesture/speech-based set of actions"*, which is in fact a possible configuration.

Figure 4 also shows the use of interaction modalities on a per-function basis. For within-page functions such as clicking, scrolling and zooming, gesture was predominant, while between-page functions such as going back and forward or switching tabs were often almost equally controlled with speech. The Click Link and Switch Tab referents were the only ones that were associated with multimodal interactions. For Click Link, this involved pointing with the hands and then clicking by saying "click", "select" or "open", while for Switch Tab, some participants suggested to say "tab" before flicking the hand left or right to navigate through tabs.

Post-Task Ratings

Figure 5 shows the average ratings of Morris's and our participants for Morris's post-study questionnaire [6]. We will first compare her results with our Task 1, before discussing differences between Tasks 1 and 2, and our observations.

In Task 1, we obtained similarly positive ratings for questions on whether participants would enjoy operating the browser in this manner, whether they had fun, whether gesture commands seemed effective, and whether the interactions felt natural. The other ratings, most notably the one that concerned the effectiveness of speech commands, were considerably lower. We could not test for statistical differences since not all ratings were reported in [6].

We also critically note that several participants first hesitated in answering some of the questions in Task 1 (e.g. P9: *"I did not operate the browser in this task."*). P10 commented: *"I just said what I want in this task so that would be how agreed or disagreed I was in above questions"*. The questions on whether they felt tired and physically uncomfortable were even deemed inappropriate by four participants (e.g. P3: *"Only proposing gestures and showing them once is not enough to feel tired or uncomfortable (at least for me)"*). These critical remarks relate to the Wizard of Oz method, specifically, the fact that there was no system dialogue.

In Task 1, participants frequently indicated a preference for gesture rather than speech. Many explained that it would be difficult to say which gesture would be better, as it was not "really possible" to see whether the gesture could actually be recognised. P25 added that *"Kinect will need to be clever"*. Then, after testing the gesture recogniser in Task 2, there was often a shift towards speech commands (P9: *"I am starting to like speech more now since it works"*). P5 commented: *"Some of the gestures felt physically uncomfortable; speech commands seemed to perform better, but I think I lack practice using gestures."* P20 criticised the lack of system dialogue especially in Task 2: *"Give more visual feedback!"*

A Friedman test to evaluate differences in the post-task ratings of Task 1 and Task 2 was significant ($p < .001$). Follow-up pairwise comparisons using Wilcoxon tests confirmed that operating the browser in this manner was significantly less enjoyable in Task 2 ($p < .03$). Also gesture commands seemed significantly less effective in Task 2 ($p < .001$). Although there were similar differences for speech commands, they were not significant. Notable is the increase in the tired and physically uncomfortable ratings in Task 2, where the former differences were significant ($p < .01$).

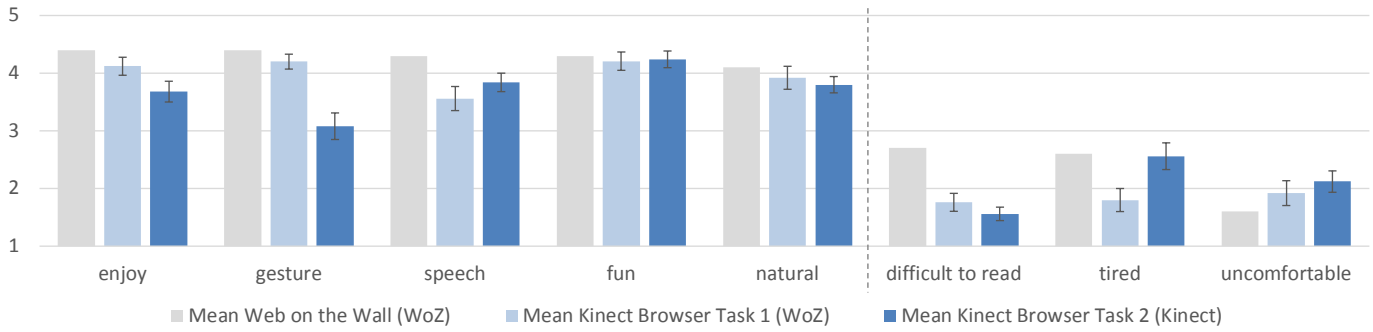

Figure 5. Mean post-study/task ratings for Morris's and our study on 5-point Likert scale (error bars show standard error)

Finally, although users were repeatedly asked to suggest gesture, speech, or multimodal interactions independent of whether the Wizard or Kinect were responsible for recognition, multimodal suggestions were more frequent in Task 1.

We note that the fixed order starting with Task 1 was required to make sure we properly replicate Morris's study. Participants were asked to fill in questionnaires and offered a break between tasks. Although unlikely, there might be some bias regarding the method and some ratings.

Feedback and Observations during Elicitation

Our detailed observations relate to pointing and clicking, use of gesture and speech, and roles of hands.

Pointing and Clicking

All participants except for one suggested using their hands for directly pointing at the screen, i.e. in x/y direction; P15 instead suggested pointing on a horizontal surface along the x/z-axes. The preferences for activating targets were less clear. Most participants felt that the click/grip gestures were most efficient and less tiring, but the many false-positives for KinectInteractions' Grip gesture caused frustration. The dwell and push gestures were generally considered problematic, as they seemed more difficult to control, more time consuming, and often accidentally clicked links. In particular, dwell caused confusion when one hand was actively used to target, while the other was already hovering a link. While the Wizard was able to distinguish which hand was actively used for pointing, our system considered both hands if they were inside the interaction window. The majority of participants expressed that it was generally difficult to select a particular target independent of which gesture was used. To avoid cursor movement during selection, multimodal interactions hand-as-mouse + "click/select/open" were preferred by 14 participants in Task 1 and 5 in Task 2.

Use of Gesture

Prior influences from touch were evident in almost all proposed gestures. Participants commonly tried to adapt touch gestures using finger movements to hand gestures using arm movements. At the same time, they frequently paid attention to potential conflicts in the suggested interaction set. For example, similar to Morris's study, some participants initially suggested to extend an arm out to the side or in front of their body and moving it up and down for the Scroll Page referent. However, they soon realised the potential for conflicts as they suggested similar interactions for the Click Link and Go Back/Forward referents. The majority then commonly suggested the grip + drag up/down gesture also implemented in Kinect Browser. Scrolling in the opposite direction to the hand movement was generally preferred and considered most similar to touch scrolling on smartphones.

Another common theme was to define mirror gestures for function pairs Scroll Up/Down, Zoom In/Out, Go Back/Forward, and Next/Previous Tab with essentially the same gesture being suggested, but then performed in opposite direction. On the other hand, this sometimes also made gesture recall more difficult and even caused conflicts. For example, similar to Morris's study, participants typically argued for either the *book* or *arrow* metaphor when suggesting flick hand gestures for going back and forward in the history or for switching tabs. Yet, when asked to demonstrate the gestures repeatedly, the metaphors were sometimes confused and participants first had to remind themselves of how flick scrolling was actually in common touch interfaces.

Roles of Hands

Most participants suggested symmetric use of hands so that they could flexibly switch between hands. Participants commonly started using the dominant hand for pointing. But they frequently switched hands, in particular, when they had to select targets on the other half of the screen. Some participants assigned opposing roles to hands. For example, P4 defined Next/Previous Tab by pointing with the right hand to the right-top corner and with the left to the left-top corner, respectively. Flick gestures were typically preferred for the Go Back/Forward rather than the Next/Previous Tab referents due to seemingly higher frequency of use. Tabs were instead frequently just clicked or controlled using speech. Still, four participants in Task 1 and 2 in Task 2 wanted to use the flick gestures for both referent pairs. To resolve conflicts, they assigned different roles to hands or used bimanual interaction. To give an example for the first case, P13 and P20 used the left hand for tabs and the right hand for history navigation. Alternatively, P3 defined flick gestures with one hand for Go Back/Forward and two hands for Next/Previous Tab.

Use of Speech

Similar to Morris's study, participants commonly used signifiers such as "go", "history" or "tab" to trigger the function pairs Go Back/Forward and Next/Previous Tab. While this was considered a natural command for the former, it felt less

natural to say "tab next/previous" vs. "next/previous tab", but still more consistent in terms of the overall command set. Yet, also speech recognition had drawbacks as it was frequently considered *"too slow"* and not all functions could equally be triggered using speech. For example, using hands for scrolling and zooming allowed for continuous scroll and zoom, while the "scroll up/down" and "zoom in/out" commands scrolled page-wise and zoomed in steps. P20 considered selecting tabs or links difficult for non-English titles.

Discussion

Overall, the feedback on Kinect Browser was quite positive. All participants could identify with using Kinect for web browsing in the studied setting, and most commonly considered it for lean-back, but also purposeful, use [5], as in our trip planning scenario. Generally, the supported interactions seemed a good fit, but not all were equally easy to perform or felt as effective. Participants particularly appreciated that Kinect Browser was ready to support most of the interactions they suggested and therefore able to adapt to their personal multimodal configurations for common browser functions.

All aspects of our system were implemented using Kinect's SDK and the KinectInteraction toolkit. Therefore, the issues noted by participants should not be attributed to poor implementation. Since our study was conducted with state-of-the-art input sensing techniques, the issues we faced and were able to study thanks to our system are likely to be faced by others. The majority of issues related to accuracy of skeletal tracking, false positives in gesture recognition (e.g. grip), and minor issues with speech recognition. As a common effect, we observed frequent switches to speech modality. P9: *"Gestures did not really work and caused a lot of frustration. I feel like I would need to concentrate more on the gestures than actually browsing."* P3: *"To perform the given tasks, I used more voice commands than previously anticipated, since they seemed to work better for me than gestures. This is also due to issues with Kinect tracking my hands (jumping dots, jittering)."* Participants attributed the issues to the Kinect sensor, our recogniser, or both (e.g. P13: *"As soon as Kinect accuracy gets better, this will be a really fun tool. Currently, the tool sadly confuses several interactions."*), but also felt that training might be required (P13: *"Like with any new system, it takes some time and exercise to get the gestures right."*).

To follow this up, we used our experiment software to produce accuracy statistics and correlated them with Kinect Browser's gesture/speech logs and our video analysis. On average, only 84% percent of processed joints were successfully tracked, and this varied by 19% between participants. Although it seemed during the study that participants frequently resorted to speech commands when skeletal tracking and gesture recognition were poor, based on the collected data, there was no such correlation. The fact that gesture was still predominant despite the gradual transition to speech was due to the fact that frequent browser functions such as Click Link required gestures and because participants often started out using gesture and, only after some time, switched to speech.

Nonetheless, pointing and clicking might be improved by adding common smoothing and filtering techniques that we have not implemented yet. For example, gesture recognition rates may be improved using more advanced relaxation techniques [16]. Other techniques from perspective-based [10], finger-based [13] and proxemic interaction [1] may also be adopted to improve the tracking based on Kinect.

While Kinect Browser can already be used with many web sites, not all requirements are fully supported. For example, text input remains an open issue, as our study was mostly concerned with within-page and between-page navigation. At this stage, bookmarks can be loaded from the host browser. In the future, our implementation could be extended with speech commands, or a specific text input interface, to enter URLs and other data in the browser or current web page, e.g. for login. We refrained from defining speech commands for all links and tabs, although this was often assumed by participants, and is also technically possible. We could automatically generate Kinect grammar once a page has been loaded, but this might drastically increase the potential for conflicts with other speech commands. It is also no general solution, as not all interactive elements are always properly labelled.

REFLECTION

We have presented our investigation around Kinect Browser, a new multimodal web browser using Kinect developed based on Morris's Web on the Wall [6] and our own study. We will use this section to reflect on our investigation. We will start with the contribution and issues of generalisability, applicability and limitations. We will then elaborate on the lessons learned. After highlighting again the main findings of our work, we will conclude with a set of guidelines and concrete recommendations for future elicitation studies.

Contribution and Limitations

There are three important parts to this research that distinguish it from prior work. First, a key aspect of our work was to develop an actual system based on Morris's guessability study as an example. While we found her study very valuable and were also successful in developing Kinect Browser, going through the process and trying to apply her findings during development was not straightforward and posed several challenges that we had not previously anticipated. Second, we wanted to investigate reproducibility by modelling Task 1 on top of Morris's study. Our community struggles with replication [2] and our work allows result comparison across papers, something that is rarely possible to this extent. Third, by deviating from the common protocol and involving an actual system in Task 2, we wanted to explore the effects of using a mixed-initiative elicitation method. Our hope with this was to first obtain an unframed interaction set, but then also expose participants to the technology for which the study was conducted and see how they would act on recognition issues and aim to resolve any conflicts concerning their proposals.

While focusing on a first example of system inclusion in the elicitation process, we believe that many of the effects we observed would generalise to other domains, other types of systems and systems with varying levels of implementation.

First, our system was designed for multimodal web browsing using Kinect. However, many of the proposed interactions

are not specific to web browsing as they concern basic tasks, such as clicking, scrolling and zooming, as well as going back and forward in the history or tabs, which are common to most applications (undo/redo, window management, etc.).

Moreover, many of the participants' proposals seemed to favour gesture-based interaction and were inspired by their previous use of touch interfaces. Similar to Morris, we would argue that, by opening our investigation to multimodal interaction using Kinect rather than limiting ourselves to touch surfaces as in previous studies [8, 14], we in turn opened the design space for a more general interaction set. Further, we showed that most proposals can be implemented using Kinect, but this is just one example of a low-cost input tracking system that we used for demonstration.

Finally, our investigation benefited from Morris's prior work since we anticipated most interaction proposals thanks to her findings. While important for the replication aspect of our paper, it might be mistaken as a limitation given that elicitation studies are usually conducted as a first step where there are no prior results. We critically note that Morris's study was not a prerequisite and that it is in our methodology that a useful basis for implementing the system is formed (cf. Task 1).

Lessons Learned

Apart from the accepted issues of elicitation studies discussed by Wobbrock et al. [14] and Morris [6], we believe that our study revealed two important issues. First, similar to Morris, Task 1 of our study was based on Wizard of Oz, relying on the experimenter rather than the system. Second, as suggested by Wobbrock et al. and commonly done by others, system dialogue was removed, but the lack of feedback was noted as an issue. We elaborate on both issues in more detail below.

Wizard of Oz was necessary to simulate parts of the system that were not implemented yet due to the lack of technical support, e.g. finger tracking using Kinect, or insufficient design knowledge, e.g. popular gesture or speech commands. However, we felt during the study that many potential sources of bias may lie in the nature of the method itself. For example, when scrolling or zooming were shown as continuous actions using the mouse wheel, participants were more likely to suggest gestures. After showing them as discrete actions using Page Up/Down and +/- keys, participants tended to suggest speech commands. Likewise, when asked for speech commands for the Reload referent, participants were likely to just repeat "reload", making it a popular suggestion. Moreover, in Task 1, participants commonly paid less attention to gesture form and potential errors, as they relied on the Wizard to react properly to their interactions. Although participants often suggested the same gestures in both tasks, our recogniser, which we had always running in the background, showed generally lower recognition rates in Task 1.

The use of Wizard of Oz for the purpose of eliciting multimodal interactions therefore also incurs a certain cost for system design as it may not be suitable to obtain accurate interaction proposals. One interesting new idea is to use a second study participant as the Wizard instead of the experimenter. Lee et al. [3] elicited freehand gestural interaction for navigating Google Earth maps using pairs of users where one would function as the *performer* and the other as the *recogniser*. Although this introduces another variable to the study design, they generally observed a good interplay of performer and recogniser. Interestingly, even human intelligence of the recogniser was not always sufficient to correctly interpret the interactions intended by the performer. This led to strong effects such as slower and larger gestures for some pairs.

The most important aspect of our investigation is the idea of dividing the elicitation process into two parts, first using a human recogniser and then a system-based recogniser. This mixed-initiative design allowed participants to first think without any technological constraints, but then also get a feel for the technology and perhaps reconsider their interaction proposals to make them feasible for implementation. To our knowledge, our study is the first to embed a system in the elicitation task. This is in contrast to Wobbrock et al. [14] and many follow-up studies that commonly argued for excluding system dialogue from the investigation in order to observe users' unrevised behaviour, and then drive system design to accommodate it. Yet, by including a system such as Kinect Browser, we were able to see interesting effects that we consider equally important and crucial to drive system design. Our lessons learned mainly manifest at three levels.

First, previous elicitation studies commonly sought consensus among participants in terms of proposed interaction sets. For referents where consensus was low, the suggested way out was to use on-screen widgets [14] or multimodal synonyms [6]. However, this is not always possible, especially not for all functions. Therefore, as also argued by others, we designed the system in a way that it can easily be adapted to support multiple alternative interactions per referent.

Second, awareness of users' *revised* behaviour, and being able to observe how they may try to work around issues, may be crucial. For example, most participants initially argued for gesture as the primary interaction modality, but gradually shifted towards speech. This demonstrated a plausible escape strategy that was only possible since we considered multimodal synonyms right from the beginning. This also echoed in significant differences between Tasks 1 and 2 in terms of whether gestures felt effective for controlling the system. It was not clear whether this change in behaviour was mostly due to poor recognition or lack of training. We then developed statistics based on our experiment software and video analysis to distinguish issues related to Kinect and our implementation. As we found no strong correlation, we might agree with those participants that felt more need for training.

Third, using the system rather than just relying on the Wizard raised the issue of designing appropriate feedback for users. Similar to Wobbrock et al. [14], our system provided no feedback to users other than showing the *effect* of their interaction as soon as it was successfully recognised. This was the same in both conditions not to increase bias towards one or the other elicitation method. Yet, this had the consequence that, in contrast to when the Wizard was responsible for recognition, there was only little feedback on Kinect's tracking accuracy. This was noted as an issue (P22: *"It would be great*

to have an indicator that shows how 'good' the Kinect sees me."). In the future, our accuracy statistics could provide live feedback on uncertainties in the tracking, allowing users to learn how they may be overcome by adjusting their pose.

Guidelines and Recommendations

Based on our experience with mixed-initiative elicitation, we offer the following guidelines and recommendations.

Draft the system: First, we find it necessary to involve a system in elicitation studies even though the key idea of obtaining a user-defined interaction set remains the same. Elicitation exercises should attempt to cover all referents required for a system to work, including basic ones such as pointing, clicking, scrolling and zooming. Clearly, the quality and accuracy of results achievable with mixed-initiative elicitation depend on the state of the implementation. While current software and hardware limitations may have a negative effect, it was particularly interesting to see how participants coped with recognition issues and explored alternative interactions.

Make it adaptable: We developed Kinect Browser in a way that it clearly separates the system's behaviour from the recognition code that triggers it. Specifically, we implemented 10 common browser functions and 25 separate interactions. This enabled us to configure the same function for different gestures and speech commands that may or may not be combined. Per configuration, Kinect Browser supports 1,585 possible mappings of interactions to browser functions so that every function is associated with at least one interaction. In principle, the recognition code can be shared, or varied, between implementations and even completely replaced to experiment with other technologies than Kinect.

Share code and data: Finally, elicitation studies should be recorded and the data shared with the community. Our review of the literature showed that this is rarely the case. But it would make results more practical and reproducible. With our custom experiment software, we developed a first set of tools for recording Kinect data and analysing user-defined interaction sets. These tools prove not only useful for our study, but could in the future also be used by others for studying and sharing their user-defined interaction sets. This may soon allow us to extract some sort of natural user interface design patterns, which our community is still relatively short of.

CODE AND DATA

To enable replication and extension of our results, our study material including the Kinect Browser source code, the multimodal interaction sets and our analysis are available at: `https://github.com/globis-ethz/kinectbrowser`.

REFERENCES

1. Ballendat, T., Marquardt, N., and Greenberg, S. Proxemic Interaction: Designing for a Proximity and Orientation-Aware Environment. In *Proc. ITS* (2010).
2. Hornbæk, K., Sander, S. S., Bargas-Avila, J. A., and Simonsen, J. G. Is Once Enough? On the Extent and Content of Replications in Human-Computer Interaction. In *Proc. CHI* (2014).
3. Lee, S.-S., Chae, J., Kim, H., Lim, Y.-K., and Lee, K.-P. Towards more Natural Digital Content Manipulation via User Freehand Gestural Interaction in a Living Room. In *Proc. UbiComp* (2013).
4. Liebling, D. J., and Morris, M. R. Kinected Browser: Depth Camera Interaction for the Web. In *Proc. ITS* (2012).
5. Lindley, S. E., Meek, S., Sellen, A., and Harper, R. H. R. Its Simply Integral to What I do: Enquiries into how the Web is Weaved into Everyday Life. In *Proc. WWW* (2012).
6. Morris, M. R. Web on the Wall: Insights from a Multimodal Interaction Elicitation Study. In *Proc. ITS* (2012).
7. Morris, M. R., and Horvitz, E. SearchTogether: An Interface for Collaborative Web Search. In *Proc. UIST* (2007).
8. Morris, M. R., Wobbrock, J. O., and Wilson, A. D. Understanding Users Preferences for Surface Gestures. In *Proc. GI* (2010).
9. Nacenta, M. A., Kamber, Y., Qiang, Y., and Kristensson, P. O. Memorability of Pre-designed and User-defined Gesture Sets. In *Proc. CHI* (2013).
10. Nacenta, M. A., Sallam, S., Champoux, B., Subramanian, S., and Gutwin, C. Perspective Cursor: Perspective-Based Interaction for Multi-Display Environments. In *Proc. CHI* (2006).
11. Oh, U., and Findlater, L. The Challenges and Potential of End-User Gesture Customization. In *Proc. CHI* (2013).
12. Ruiz, J., Li, Y., and Lank, E. User-Defined Motion Gestures for Mobile Interaction. In *Proc. CHI* (2011).
13. Vogel, D., and Balakrishnan, R. Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays. In *Proc. UIST* (2005).
14. Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-Defined Gestures for Surface Computing. In *Proc. CHI* (2009).
15. Wobbrock, J. O., Wilson, A. D., and Li, Y. Gestures without Libraries, Toolkits or Training: A $1 Recognizer for User Interface Prototypes. In *Proc. UIST* (2007).
16. Wu, M., Shen, C., Ryall, K., Forlines, C., and Balakrishnan, R. Gesture Registration, Relaxation, and Reuse for Multi-Point Direct-Touch Surfaces. In *Proc. Tabletop* (2006).